June 1976

Tests without power. A rejoinder. *)

Erling Sverdrup
University of Oslo

Abstract This is a rejoinder to Per Martin-Löf's (2) reply to my comments on his paper on the notion of redundancy (1).

1. The main problem

Let me recall what the problem is about. Per Martin-Löf (1) has stated (page 4): "I regard it as a fundamental principle that the smaller the number $f(t(x))$ of outcomes is that realize the observed value of $t(x)$, the more does our observation x contradict the hypothesis that the statistic $t(x)$ can be reduced to $u(x) = u(t(x))$. By fundamental principle, I mean it does not seem possible to reduce it to any other more basic or convincing principles". This is, of course, a very provocative statement. It contradicts the Neyman-Pearson point of view, which many statisticians have considered convincing and fundamental; and it contradicts the likelihood idea, which some statisticians consider fundamental. Repudiating the mainstream of ideas in statistical inference theory, Per Martin-Löf should have expected a critical appraisal of his method. It would have been natural if he had undertaken to do so himself. At least he should have left open the possibility of doubt. My examples show that his principle is not so convincing and they lend support to the now classical

*) This is a new version of a previously published paper with the same title.

ideas in inference theory. All of the methods suggested in my examples could have been derived from the general results by Neyman and Pearson (extended to the discrete case) in their famous 1933 paper. (In example 1 a specification of the alternative is needed.)

Per Martin-Löf reiterates his idea in a strengthened form in his reply, where he states: "I know of no case when one has rejected a statistical hypothesis after having observed the most probable value under the hypothesis, and I would not do so myself". This is a rather spectacular statement. Statisticians working with the likelihood principle or the Neyman-Pearson ideas have been aware of this implication and they have had no misgivings for that reason. I am not certain if Martin-Löf's statement is meant to be general. If so, there exists an abundance of examples showing that practical statisticians would reject when the most probable event occurs. If his statement is meant to be subject to his assumptions about uniform conditional distribution and confinement to 'relevant' statistics, then Per Martin-Löf should explain why it is not unreasonable to reject when the most probable event occurs and these assumptions are not fulfilled whereas it is unreasonable when they are fulfilled. In any case my examples serve their purpose.

In my example 1 it is obvious from a practical statistical point of view that it is the hands that are adventageous to the dealer that are conspicious and under the hypothesis there need not be any connection between the advantageousness of a hand and the probability of its occurance. In example 4 when testing $\sigma_1 = \sigma_2$ in the case of two samples from normal populations, we would certainly reject if the observations in the first sample are very close together whereas they are scattered in the second sample. Again,

this contradicts Martin-Löf's principle, and rejects "after having observed the most probable outcome under the hypothesis". Statisticians have used the F-test through the last 45 years and they have never warned against using it when the numbers of degrees of freedom are small. Thus there certainly has been "agreement among statisticians" at least in this case (see Martin-Löf (2) p. 165).

In this connection let me also emphasize that statistical inference theory has gone a long way since its modern development started at the beginning of this century. Milestones have been left behind and results of lasting value and immeasurable importance have been established. Of course they go far beyond confirmations of what have been common practice among statisticians. The present journal - SJS - has a mission in furthering the right understanding of what inference theory stands for today. Of course, much needs to be done, but the results obtained should constitute a sound foundation for development of new ideas.

2. The concept of reductiveness

Martin-Löf's definition of this concept is found on page 4 and is admittedly a meaningful and interesting concept from a statistical point of view. (It really is a condition for obtaining a test with "Neyman-structure", according to Lehmann's terminology.) However,

I do not understand Martin-Löf's (2) remark that "if we consider one-sided alternatives $p_A > p_B$, the hypothesis is not reductive and hence falls outside of my framework". Now, perhaps it is not necessary to discuss this matter since all my examples, that are not already "two-sided", could be made so. However, I am a little curious. It seems to me that in example 3 (for instance) $t = (X, Y_1 + Y_2)$ is sufficient if $p_A \geq p_B$, whereas $u = X + Y_1 + Y_2$ is sufficient if $p_A = p_B$. Hence u is a function of t, with no unique inverse. I had no difficulties in interpreting Martin-Löf's idea about reductiveness in his first paper. If $t(x)$ is "relevant" (minimal sufficient) a priori and $u(x)$ relevant under the hypothesis, then $u(x)$ should be a function of $t(x)$ but not vice verca. This definition is satisfactory in the discrete case and could, in a modified form, be made satisfactory in the general case (introducing the sigmafields generated by $t(x)$ and $u(x)$). Therefore, Martin-Löf's remarks about reductiveness in his second paper came as a surprise to me. He should certainly explain why a hypothesis $\theta = 0$ is not reductive when tested against $\theta > 0$ whereas it is reductive when tested against $\theta \neq 0$. It would be of interest to know his explicite definition of reductiveness.

3. One-sided and two-sided alternatives

First, I do certainly not find it contradictary to use a "one-sided" test in a "two-sided" test situation. After all, it is too easy to find examples where such is the case. I have made no statement which Martin-Löf could interprete in such a manner.

The regirement of "two-sided-ness" is hard for me to understand. I would like to comment on that even if

I may then do Martin-Löf some injustice. He may have something else in mind.

First a trivial remark. The normal distribution is characterized by the mean μ and the variance σ^2. It could just as well have been characterized by the second order moment $\sigma^2 + \mu^2$ and the inverse coefficient of variation μ/σ. In general any parameter characterization θ could be replaced by a parameter characterization $\theta' = f(\theta)$ if f is one-to-one between the range of θ and the range of θ'. Let now $\theta = (\mu, \sigma)$ where μ is any scalar parameter subject to testing, whereas σ is a nuisance parameter (which may be many-dimensional or absent altogether). Now if the problem is to test $\mu = 0$ against $\mu \neq 0$ it is easy to see that we could replace $\theta = (\mu, \sigma)$ by $\theta' = (\mu', \sigma)$ where the problem now is one-sided and the relation between θ and θ' is one-to-one, $\mu = 0$ corresponds to $\mu' = 0$ and $\mu \neq 0$ corresponds to $\mu' > 0$.

Let $\mu' = \frac{\mu}{1+\mu}$ if $\mu \geq 0$ and $\mu' = 1 - \frac{1}{\mu}$ if $\mu < 0$. Another way of proceeding is to let $\theta' = (\mu', \sigma')$ where $\mu' = |\mu|$ and $\sigma' = (\delta, \sigma)$, $\delta = +1$ or -1 according as $\mu > 0$ or < 0. Then $\sigma' = (\delta, \sigma)$ would be the new nuisance parameter and the problem would be to test $\mu' = 0$ against $\mu' > 0$. Hence, any of Martin-Löf's own examples could be made one-sided (and thus not be reductive?) Vice versa, any of my "one-sided" problems could be made two-sided (and hence reductive?)

My own use of the words "one-sided" and "two-sided" is only descriptive in connection with the special choice of parameters in the examples. (In the original version of my paper I did not need those expressions. It was the editor and referee who urged me to consider "two-sided" situations.) The situation is somewhat different if

two-sided-ness is connected with a requirement concerning power and "distance" from hypothesis. However, still it is difficult to see the connection with "reductiveness".

4. Equal tails

Of course, in example 4 I could have determined $c_1(W)$ and $c_2(W)$ such that $\Pr(z_1 < c_1(W)) = \varepsilon_1 \neq \frac{\varepsilon}{2}$ and $\Pr(z_1 > c_2(W)) = \varepsilon_2 \neq \frac{\varepsilon}{2}$, where $\varepsilon_1 + \varepsilon_2 = \varepsilon$, and, of course, the choice of ε_1 and ε_2 is "arbitrary" just as the choice of the level ε itself is "arbitrary". I would not choose $\varepsilon_1 = 0$ (or $\varepsilon_2 = 0$), since it would impair the sensitivity (power) relatively to alternatives $\sigma_1 < \sigma_2$ (or $\sigma_1 > \sigma_2$). I would not choose $\varepsilon = 0$ either, since it would impair the sensitivity altogether.

Since the problem under discussion was an entirely different one, I did not want to go into the choice of $c_1(W)$ and $c_2(W)$. I just followed what has been commonly accepted in textbooks, see e.g. Anders Hald, page 380 or M.G. Kendall, page 115-116. From a practical point of view the problem is not very important. However, it could be used to illustrate some decision-theoretical aspects. Thus in my paper (2) (page and) I have pointed out that in the natural three-decision problem in this situation, $\varepsilon_1 = \varepsilon_2 = \frac{\varepsilon}{2}$ leads to a test which uniformly maximizes the performance among all performance unbiased tests. Similar optimum properties could be formulated for any choice of ε_1 and ε_2. That the distribution under the hypothesis lacks symmetry is irrelevant. Thus, to a great extent I am willing to go along with Martin-Löf when he says that the unbiasedness in power must be rejected in situations with two-sided alternatives. I have not advocated its use. In the case of

one-sided alternatives the principle of power unbiasedness is of course very important, and needed.

5. The n-dimensional Riemann-surface

My example 4 is treated differently by Martin-Löf and myself. Both treatments are in accordance with Martin-Löf's prescription in (1). Suppose that S^2 is the estimate of σ^2 based on three normal observations. Then the difference between his and my treatment is analogous to using S and $S^2 = Z$ respectively. The densities are respectively const. $s\, e^{-\frac{s^2}{\sigma^2}}$ and const. $e^{-\frac{z}{\sigma^2}}$. Hence most "probable" value of S is $\sigma/\sqrt{2}$, whereas the most probable value of S^2 is 0. In (2) he introduces the assumption that the "relevant" statistic should be a "metric", thus excluding S^2. I do not see the statistical reasoning behind requiring that the relevant statistic shall have the form of a metric like S.

Of course, the points taken up in sections 3 - 5 above are very minor points. The important ideas are discussed in sections 1 and 2.

References

Hald, Anders. Statistical Theory with Engineering Applications. New York 1952.

Kendall, Maurice G. The Advanced Theory of Statistics, Vol II. London 1946.

(1) Martin-Löf, Per (1974). The Notion of Redundancy and its use as a Quantitative Measure of Discrepancy between a Statistical Hypothesis and a set of Observational Data, Scand. J. Statist. 1, 3-18.

(2) Martin-Löf, Per (1975). Reply to Sverdrup's Polemic Article Tests without Power, 161-165. Scand. J. Statist.

Neyman, J. and Pearson E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. 231, 289-337.

(1) Sverdrup, Erling (1975). Tests without Power. Scand. J. Statist. 2, 158-160.

(2) Sverdrup, Erling (1976). Significance testing in multiple statistical inference. Scand. J. Statist., 3